Journal of Multivariate Analysis 100 (2009) 670–685

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

On weighting of bivariate margins in pairwise likelihood

Harry Joe [a], Youngjo Lee [b,*]

[a] University of British Columbia, Department of Statistics, Canada

[b] Seoul National University, Department of Statistics, Republic of Korea

ARTICLE INFO

Article history:
Received 4 January 2008
Available online 17 July 2008

AMS 2000 subject classifications:
primary 62H12
secondary 62F12

Keywords:
Composite likelihood
Binary probit
Clustered data
Longitudinal data

ABSTRACT

Composite and pairwise likelihood methods have recently been increasingly used. For clustered data with varying cluster sizes, we study asymptotic relative efficiencies for various weighted pairwise likelihoods, with weight being a function of cluster size. For longitudinal data, we also study weighted pairwise likelihoods with weights that can depend on lag. Good choice of weights are needed to avoid the undesirable behavior of estimators with low efficiency. Some analytic results are obtained using the multivariate normal distribution. For clustered data, a practically good choice of weight is obtained after study of relative efficiencies for an exchangeable multivariate normal model; they are different from weights that had previously been suggested. For longitudinal data, there are advantages to only include bivariate margins of adjacent or nearly adjacent pairs in the weighted pairwise likelihood.

1. Introduction

Composite likelihood methods based on optimizing sums of log-likelihoods of low-dimensional margins have been considered by many authors in recent years; they are useful for multivariate models in which the likelihood of multivariate data is too time-consuming to compute. In particular, pairwise likelihood or bivariate composite likelihood methods are based on bivariate margins. An excellent review paper on composite likelihood is [21]. Other recent references include: [7,17,10,14,1,19,2,26,23–25,22]. The term *composite likelihood* originates from [15]. Composite likelihood methods have been applied for multivariate probit and other models for correlated binary and ordinal response data, binary spatial data, copula and mixture models for count data, etc.

Suppose that there are n experimental units (or clusters), and $d_i \geq 2$ observations or repeated measurements for the ith unit. The data are vectors $\mathbf{Y}_i = (Y_{1i}, \ldots, Y_{id_i})$, $i = 1, \ldots, n$. The index i stands for a cluster/family for clustered/familial data, and a subject for longitudinal data. Let $f_{\mathbf{Y}_i}(\cdot; \boldsymbol{\theta})$ be the joint density of a parametric model for the data, with parameter vector $\boldsymbol{\theta}$. Let $f_{Y_{ij},Y_{ik}}(\cdot; \boldsymbol{\theta})$ denote the bivariate marginal density for the (j, k) margin of the ith unit.

The pairwise or bivariate composite log-likelihood (BCL) has form

$$L_w = \sum_i \sum_{j<k} w_{i,jk} \log f_{Y_{ij},Y_{ik}}(y_{ij}, y_{ik}; \boldsymbol{\theta}), \tag{1.1}$$

where the $w_{i,jk}$ are weights. (We use the abbreviation BCL for bivariate composite log-likelihood, because the abbreviation PL is sometimes used for pairwise likelihood, pseudo-likelihood, penalized likelihood or partial likelihood.) When $w_{i,jk} \equiv 1$, this is called the unweighted BCL. The estimator $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ that maximizes (1.1) is called the BCL estimator. Under regularity conditions, this is the same as the (unique) solution of $\mathbf{g} = n^{-1}\partial L_w/\partial\boldsymbol{\theta} = 0$. For the pairwise likelihood approach to work, the parameter vector $\boldsymbol{\theta}$ should be identifiable from the set of bivariate margins.

* Corresponding author.
E-mail addresses: harry@stat.ubc.ca (H. Joe), youngjo@snu.ac.kr (Y. Lee).

0047-259X/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2008.07.004

Using the theory of estimating equations, the asymptotic covariance matrix of $n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is $\mathbf{V} = \mathbf{D}^{-1}\mathbf{M}(\mathbf{D}')^{-1}$, where $\mathbf{D} = \lim_{n\to\infty} \mathrm{E}\,[-\partial \mathbf{g}/\partial \boldsymbol{\theta}]$ and $\mathbf{M} = \lim_{n\to\infty} n\mathrm{Cov}\,(\mathbf{g})$. For asymptotic relative efficiencies (AREs), we take ratios of diagonal elements of the inverse Fisher information matrix (the asymptotic variance matrix of the maximum likelihood estimate) with the corresponding diagonal elements of $\mathbf{V}$.

For clustered data such that $\mathbf{Y}_i$ has dependence structure close to exchangeable or some familial dependence pattern (with sib-sib correlation, parent-offspring, degree 2 relation etc.), we consider weights $w_{i,jk} = w_i$ that are functions of the cluster size d_i and independent of the members of the cluster. If all clusters have the same size, then the discussion of w_i is not needed; in this case, one can take $w_i = 1$ without loss of generality. For longitudinal data, the weights $w_{i,jk}$ in general could depend on the time lag $|k - j|$.

In this paper, we study the weighting of BCLs for better ARE in various situations. A weighted pairwise log-likelihood can be the log-likelihood in some situations, retaining the full efficiency. However, sometimes the ARE can be poor; we indicate situations when it happens. Three special cases are given below for which (1.1) is the actual log-likelihood for appropriate choices of weights.

1. If observations are independent within each cluster, so that $f_{\mathbf{Y}_i} = \prod_{j=1}^{d_i} f_{Y_{ij}}$, then the choice $w_{i,jk} = 1/(d_i - 1)$ in (1.1) results in the log-likelihood.
2. For discrete observations such that identical univariate margins and perfect dependence hold, then for all clusters $y_{i1} = \cdots = y_{id_i}$, $f_{\mathbf{Y}_i}(\mathbf{y}_i) = f_{Y_{i1}}(y_{i1})$ for $\mathbf{y}_i = y_{i1}\mathbf{1}$, and $f_{Y_{ij},Y_{ik}}(y_{ij}, y_{ik}) = f_{Y_{i1}}(y_{i1})$ for $y_{ij} = y_{ik} = y_{i1}$. The choice $w_{i,jk} = 1/[d_i(d_i - 1)]$ in (1.1) results in the log-likelihood.
3. For longitudinal Gaussian data based on AR(1) time series, the log-likelihood has the form

$$\sum_i \left[\log f_{Y_{i1}}(y_{i1}; \boldsymbol{\alpha}) + \sum_{j=2}^{d_i} f_{Y_{ij}|Y_{i,j-1}}(y_{ij}|y_{i,j-1}; \boldsymbol{\theta}) \right] = \sum_i \left[-\sum_{j=2}^{d_i-1} \log f_{Y_{ij}}(y_{ij}; \boldsymbol{\alpha}) + \sum_{j=2}^{d_i} f_{Y_{i,j-1},Y_{ij}}(y_{i,j-1}, y_{ij}; \boldsymbol{\theta}) \right], \quad (1.2)$$

where $\boldsymbol{\alpha}$ consists of elements of $\boldsymbol{\theta}$ parametrizing univariate parameters. If the univariate parameters are assumed known, and only dependence parameters are estimated, then maximizing (1.2) is the same as maximizing (1.1) with $w_{i,jk} = 1$ if $k = j + 1$ and 0 otherwise.

Cases 1 and 2 mean that the optimal weights should depend on where is the information within clusters; the strength and nature of dependence contribute to the information. This will be shown with some examples in Section 2. Some comparisons of variations of the BCL for case 3 are given in Section 3.

For clustered data, Le Cessie and Van Houwelingen [11] and Zhao and Joe [26] used the boundary case of independence to suggest that clusters be inversely weighted by a factor $(d_i - 1)$. With constant weights over varying cluster sizes, observations in the large clusters are given more weight than those in the small clusters, whereas they should be treated equally under independence. Kuk and Nott [10] agreed with the use of the weighted pairwise likelihood, with factor $1/(d_i - 1)$, for inference for univariate regression parameters, but suggest the unweighted pairwise likelihood for inference about association parameters. Geys et al. [4] have the same conclusion based upon arguments using estimating equations. Renard et al. [19] generally supported this conclusion, but numerically found that no method is uniformly better than the other.

We study several weights for BCL estimators and found that different weights are better for different parameters. Also, the recommended $w_i = 1/(d_i - 1)$ could give very low efficiency in estimating the mean parameter in highly unbalanced one-way random-effect models. Weights that are midway between the weights corresponding to independence and perfect dependence can be recommended for a generally good performance over a range of dependence. Details are given in Section 2.

In Section 3, we study several weighting schemes for AR(1) models with weights depending on lag. There can be differences in behavior of the ARE compared with random effect models. In the estimation of the location parameters in autogressive AR(1) models for longitudinal data; the ARE of the BCL estimator of some parameters could tend to one as d increases. For clustered data, typically the ARE decreases as the cluster size increases.

In non-normal models it is hard to find appropriate weights because explicit forms of AREs are rarely available, so that our approach is to study the weights of normal models analytically and then apply them to similar non-normal models, such as multivariate probit models or probit auto-regressive models. The ARE analyses for normal models are useful to understand those for similar probit models.

2. Clustered data

With clustered data, we use $w_{i,jk} = w_i$ in (1.1) with weight w_i being a function of the cluster size d_i. We assume $d_i \geq 2$ for all i. For illustration of the theory, we use models that assume exchangeability within clusters. The simplest choice is the exchangeable multivariate normal (or random effect) model — this was used by Cox and Reid [1] for the case of clusters of constant size $d_i = d$, so that they considered the unweighted BCL with $w_{i,jk} = 1$.

For the exchangeable d-variate normal distribution, the mean vector and covariance matrix are respectively

$$\mu\mathbf{1}_d \quad \text{and} \quad \Sigma_d = \eta^2\mathbf{R}(\rho), \quad (2.1)$$

where $\mathbf{R}(\rho) = [(1-\rho)\mathbf{I}_d + \rho\mathbf{J}_d]$, $\mathbf{I}_d$ is the identity matrix of order p and $\mathbf{J}_d$ is the $d \times d$ matrix of 1s. The univariate parameters are the mean μ and variance η^2, and the dependence or correlation parameter is ρ.

Table 1
ARE of BCL estimators for exchangeable probit; true parameters $\mu = -0.84$, $\rho = 0.9$

d	$\widehat{\mu}_u$	$\widehat{\rho}_u$
2	1.000	1.000
3	0.993	0.997
5	0.979	0.956
7	0.969	0.914
9	0.962	0.879
11	0.955	0.850
13	0.951	0.826
15	0.946	0.806

2.1. Exchangeable probit model with constant cluster size

Cox and Reid [1] found that the ARE of the BCL estimator for ρ, with η^2 known, is generally high and decreases as d increases. We show that this holds for the exchangeable probit model, which has two parameters, a latent mean μ and a latent correlation ρ. For dimension d, the stochastic representation is:

$$Y_j = I(Z_j \le \mu), \quad j = 1, \ldots, d, \qquad (Z_1, \ldots, Z_d)' \sim N(\mathbf{0}, \mathbf{R}(\rho)).$$

With covariates, and with more general familial correlations, this is a model used in [26], and is an example of a model for which maximum likelihood estimation is too time consuming for large cluster/family sizes.

Because of the common cluster size $d_i = d$ we study the unweighted BCL estimates $\widehat{\mu}_u$ and $\widehat{\rho}_u$. Numerical calculations of AREs are based on functions in the R package *mprobit* (http://www.r-project.org). Table 1 shows that the ARE decreasing as dimension d increases when the cluster size is fixed. The ARE is worse for larger values of ρ and Table 1 uses $\rho = 0.9$ with $\mu = -0.84 = \Phi^{-1}(0.2)$, where Φ is the standard normal cumulative distribution function. Not surprisingly there is more loss of efficiency in higher dimensions if inference is just based on bivariate margins. However, the decrease is slow as the dimension increases. In [16], the same pattern was seen for an item response model with the number of parameters equal to twice the dimension. For cluster sizes commonly seen in data, the efficiency is good.

2.2. One-way random-effect model with varying cluster sizes

We investigate the performance of the BCL estimate for models with varying cluster sizes. For this purpose we study the unbalanced one-way random-effect model or (2.1) with $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{id_i})' \sim N(\mu \mathbf{1}_{d_i}, \Sigma_{d_i})$, $i = 1, \ldots, n$. For this model, maximum likelihood estimation is numerically feasible, but we can also easily get some analytical expressions to compare the ARE of different choices of w_i that are functions of d_i. With varying cluster sizes the unweighted BCL can have very low ARE, so that the proper choice of weights is important. However, we show that the often recommended weight $1/(d_i - 1)$ could also be as bad in achieving a low ARE in part of the parameter space.

When all three parameters are estimated, the discussion of efficiency loss for weighted BCL is only relevant when the cluster size is not constant. When $d_i = d$ for all n, it can be shown that the BCL estimate is exactly the same as the maximum likelihood estimator (MLE): $\widehat{\mu} = (nd)^{-1} \sum_i \sum_j Y_{ij}$, $\widehat{\eta}^2 = \overline{S}_1/d$, $\widehat{\rho} = (d-1)^{-1}[\overline{S}_2/\overline{S}_1 - 1]$, where $\overline{S}_1 = n^{-1} \sum_i \sum_j (Y_{ij} - \widehat{\mu})^2$, $\overline{S}_2 = n^{-1} \sum_i [\sum_j (Y_{ij} - \widehat{\mu})]^2$. With fixed cluster size and η^2 assumed known, the BCL estimator of ρ is not the same as the MLE; [1] have results on the ARE of the BCL estimator of ρ in this case.

Let $Y_{i+} = \sum_{j=1}^{d_i} Y_{ij}$. Using results about the exchangeable multivariate normal distribution (given in Appendix A.1), the negative log-likelihood is:

$$\begin{aligned} -L_0 = & \sum_i \left\{ \frac{1}{2} d_i \log \eta^2 + \frac{1}{2}(d_i - 1)\log(1-\rho) + \frac{1}{2}\log[1 + (d_i - 1)\rho] \right\} \\ & + \frac{1}{2\eta^2(1-\rho)} \sum_i \left[\sum_{j=1}^{d_i} (y_{ij} - \mu)^2 - \frac{\rho}{1 + (d_i - 1)\rho} (y_{i+} - d_i \mu)^2 \right], \end{aligned} \tag{2.2}$$

and the negative weighted BCL is

$$\begin{aligned} -L_1 = & \sum_i w_i \left\{ \binom{d_i}{2} \log \eta^2 + \frac{1}{4} d_i (d_i - 1) \log(1 - \rho^2) \right\} \\ & + \frac{1}{2\eta^2(1-\rho)} \sum_i w_i \sum_{1 \le j < k \le d_i} \left[(y_{ij} - \mu)^2 + (y_{ik} - \mu)^2 - \frac{\rho}{(1+\rho)} (y_{ij} + y_{ik} - 2\mu)^2 \right]. \end{aligned} \tag{2.3}$$

The latter Eq. (2.3) uses the identity $(1 - \rho^2)^{-1}[z_1 + z_2 - 2\rho z_1 z_2] = (1-\rho)^{-1}[z_1 + z_2 - \rho(z_1 + z_2)^2/(1+\rho)]$.

To see the patterns in weights w_i that are "practically good", we do some analysis with one parameter from μ, η^2, ρ to be estimated assuming the others are known. We start with the case of estimating μ with ρ known. This analysis will suggest a good compromise choice of w_i for general use.

Varying cluster size; estimation of μ with ρ known

From solving $\partial L_1/\partial\mu = 0$, the BCL estimator is

$$\widehat{\mu}_w = \frac{\sum_i w_i(d_i-1)Y_{i+}}{\sum_i w_i(d_i-1)d_i}. \tag{2.4}$$

When $w_i = (d_i-1)^{-1}$, then $\widehat{\mu}_w = \sum_i Y_{i+}/\sum_i d_i$, the overall sample mean.

From solving $\partial L_0/\partial\mu = 0$, the MLE (with ρ known) is

$$\widehat{\mu} = \frac{\sum_i [1+(d_i-1)\rho]^{-1}Y_{i+}}{\sum_i [1+(d_i-1)\rho]^{-1}d_i}, \tag{2.5}$$

and this is not the sample mean unless d_i is constant. If ρ is unknown, then the MLE of μ involves an estimated ρ.

From comparing (2.5) with (2.4), the optimal weight w_i depends on the cluster size d_i and the correlation ρ. For the BCL, if ρ were known, the optimal w_i is

$$w_i = (d_i-1)^{-1}[1+(d_i-1)\rho]^{-1}. \tag{2.6}$$

Below we consider some weights that do not depend on ρ:

(a) $w_i = 1$;
(b) $w_i = (d_i-1)^{-1}$;
(c) $w_i = (d_i-1)^{-1}[1+\frac{1}{2}(d_i-1)]^{-1}$ (from substituting $\rho = \frac{1}{2}$ in (2.6));
(d) $w_i = (d_i-1)^{-1}d_i^{-1}$.

The ideas here, as mentioned in Section 1, are that (i) $w_i = (d_i-1)^{-1}$ is the weight such that the BCL is a log-likelihood for the case of independence; (ii) $w_i = [d_i(d_i-1)]^{-1}$ is the correct weight to use in the case of perfect dependence since in this case the information for any pair is the same.

The variance of is Y_{i+} is $\eta^2 d_i[1+(d_i-1)\rho]$, so that

$$\text{Var}\,(\widehat{\mu}_w) = \eta^2 \frac{\sum_i w_i^2(d_i-1)^2 d_i[1+(d_i-1)\rho]}{\left\{\sum_i w_i(d_i-1)d_i\right\}^2}. \tag{2.7}$$

For cases (a)–(d), let V_a, V_b, V_c, V_d respectively denote the variance in (2.7). For comparison, the variance of MLE $\widehat{\mu}$, with ρ known, is

$$V_{\text{MLE}} = \frac{\sum_i \eta^2 d_i/[1+(d_i-1)\rho]}{\left\{\sum_i d_i/[1+(d_i-1)\rho]\right\}^2} = \eta^2\left\{\sum_i d_i/[1+(d_i-1)\rho]\right\}^{-1}. \tag{2.8}$$

These are finite sample variances.

Note that $V_a \geq V_b$ always. To establish this, let $\delta_i = d_i/d_+$, where $d_+ = \sum d_i$, and let $\alpha_i = d_i - 1$ and $\beta_i = 1+(d_i-1)\rho$. Then $V_a \geq V_b$ follows from

$$\sum_i \delta_i\alpha_i^2\beta_i \geq \sum_i \delta_i\beta_i \cdot \sum_i \delta_i\alpha_i^2 \geq \sum_i \delta_i\beta_i \cdot \left(\sum_i \delta_i\alpha_i\right)^2$$

or

$$\frac{\sum_i d_i(d_i-1)^2[1+(d_i-1)\rho]}{\left\{\sum_i d_i(d_i-1)\right\}^2} \geq \frac{\sum_i d_i[1+(d_i-1)\rho]}{\left\{\sum_i d_i\right\}^2}.$$

That is, the unweighted BCL estimator of μ with varying cluster size is dominated based on variance. Otherwise V_b, V_c, V_d are never uniformly dominant over all $0 \leq \rho < 1$ and choices of $\{d_i\}$.

For (a)–(d), consider the ratios $RE_h = V_{\mathrm{MLE}}/V_h$, for $h \in \{a, b, c, d\}$. To study some extreme cases of RE_h, we specialize (2.7) and (2.8) to the case where there are cluster sizes d_1 and d_2 with frequencies m_1, m_2 respectively. The weights for cluster sizes d_1, d_2 are w_1 and w_2 respectively. Then we obtain:

$$\mathrm{Var}\,(\widehat{\mu}_w) = \eta^2 \frac{m_1 w_1^2 (d_1-1)^2 d_1[1+(d_1-1)\rho] + m_2 w_2^2 (d_2-1)^2 d_2[1+(d_2-1)\rho]}{\{m_1 w_1 (d_1-1) d_1 + m_2 w_2 (d_2-1) d_2\}^2},$$
$$V_{\mathrm{MLE}} = \eta^2 \{m_1 d_1/[1+(d_1-1)\rho] + m_2 d_2/[1+(d_2-1)\rho]\}^{-1}.$$

Some properties, partly based on numerical results, are the following.

1. For weights (b), $w_i = (d_i-1)^{-1}$, $i = 1, 2$, $RE_b \to 0$ for all $0 < \rho < 1$ for d_1 fixed, $m_2 = 1$ and $m_1 = d_2 = k \uparrow \infty$:
$$\begin{aligned} RE_b &= \frac{1}{\frac{kd_1}{1+(d_1-1)\rho} + \frac{k}{1+(k-1)\rho}} \Big/ \frac{kd_1[1+(d_1-1)\rho] + k[1+(k-1)\rho]}{(kd_1+k)^2} \\ &\sim \frac{1}{\frac{kd_1}{1+(d_1-1)\rho} + \frac{1}{\rho}} \Big/ \frac{\rho}{(d_1+1)^2} \to 0. \end{aligned}$$
2. For weights (c) or (d), that is, $w_i = (d_i-1)^{-1}[1+0.5(d_i-1)]^{-1}$ or $w_i = (d_i-1)^{-1} d_i^{-1}$, $i = 1, 2$, RE_c and RE_d can not get too low unless ρ is near zero. For example, for $\rho \geq 0.2$, $RE_c \geq 0.75$ and $RE_d \geq 0.75$. The estimators with weights (c) and (d) do not depend on ρ, and we would comparing the estimator $\widehat{\mu}_w$ to the MLE of μ with ρ known. The AREs of $\widehat{\mu}_w$ with respect to the MLE of μ with ρ estimated are higher.

The above discussion shows that the use of weights $w_i = (d_i-1)^{-1}$ for the BCL can lead to inefficient estimators when there is one large cluster. Note that the best choice of weights as a function of cluster size depends on the amount of dependence. In the subsequent analysis below, we show that the optimal weight is not the same for each parameter (in a multi-parameter family).

Varying cluster size; estimation of ρ with μ, η^2 known.

Let y_{ij} be the jth observation in the ith cluster, and $z_{ij} = (y_{ij}-\mu)/\eta$, and let $\mathbf{z}_i = (z_{i1}, \ldots, z_{id_i})'$. The negative BCL (2.3) in ρ can be written as:

$$-L_1(\rho) = \sum_i w_i \left\{ \binom{d_i}{2} \log \eta^2 + \frac{1}{4} d_i(d_i-1)\log(1-\rho^2) + \frac{1}{2(1-\rho^2)} \sum_i \left[(d_i-1)\mathbf{z}_i'\mathbf{z}_i - \rho \mathbf{z}_i' \mathbf{A}_i \mathbf{z}_i \right] \right\}, \tag{2.9}$$

where $\mathbf{A}_i = \mathbf{J}_{d_i} - \mathbf{I}_{d_i}$. Let $\widehat{\rho}_w$ be the solution to $\partial L_1/\partial\rho = 0$. Also, let

$$u_i = d_i^2\rho^2(1-\rho)^2 + d_i(2\rho - 3\rho^2 + 8\rho^3 - 3\rho^4) + (1-\rho)^2(1+3\rho^2).$$

The asymptotic variance (as $n \to \infty$) of the BCL estimator $\widehat{\rho}_w$ is $n^{-1}V = M/D^2$ where D, M are given in (A.1) and (A.2) respectively. With some algebra,

$$n^{-1}V = \frac{2(1-\rho)^2}{(1+\rho^2)^2} \times \frac{\sum_i w_i^2 d_i(d_i-1)u_i}{\left[\sum_i w_i d_i(d_i-1)\right]^2}.$$

From the Cauchy–Schwarz inequality, V is minimized when $w_i \propto u_i^{-1}$. For $\rho = 0, 0.5, 1$, the optimal weight is:

$$w_i \propto \begin{cases} 1, & \rho = 0; \\ [0.0625 d_i^2 + 1.0625 d_i + 0.4375]^{-1}, & \rho = 0.5; \\ d_i^{-1}, & \rho = 1. \end{cases} \tag{2.10}$$

These are different in form than those for the estimation of μ. The first case shows that unweighted BCL is best for weak dependence. For the two cases in (2.10) for moderate to strong dependence, the inverse weight is close to linear for small d_i (the coefficient 0.0625 of quadratic term is small), that is, as can be seen in a plot, both are close to the previously considered weight $(d_i-1)^{-1}$ for small d_i.

Varying cluster size, estimation of η^2 with ρ, μ known.

Let $\mathbf{B}_i = (d_i-1)\mathbf{I}_{d_i} - \rho\mathbf{A}_i = (d_i - 1 + \rho)\mathbf{I}_{d_i} - \rho\mathbf{J}_{d_i}$. From the above, the negative BCL (2.3) in η^2 is:

$$-L_1(\eta^2) = \sum_i w_i \left\{ \binom{d_i}{2} \log \eta^2 + \frac{1}{4} d_i(d_i-1)\log(1-\rho^2) + \frac{1}{2\eta^2(1-\rho^2)} \sum_i \mathbf{z}_i' \mathbf{B}_i \mathbf{z}_i \right\}.$$

Table 2
Multivariate exchangeable normal: ARE of BCL estimates of μ, η^2 and ρ for different distributions of varying cluster sizes; AREs are invariant to μ, η^2

ρ	Mixture	wt. in (a)			wt. in (b)			wt. in (c)			wt. in (d)		
		μ	η^2	ρ	μ	η^2	ρ	μ	η^2	ρ	μ	η^2	ρ
0.2	0.5 for $d=3,4$	0.938	0.958	0.983	0.996	0.999	0.994	0.998	0.991	0.966	0.994	0.984	0.954
	0.5 for $d=3,6$	0.795	0.862	0.923	0.972	0.996	0.982	0.990	0.949	0.861	0.972	0.917	0.816
	0.5 for $d=3,8$	0.733	0.824	0.886	0.944	0.990	0.979	0.982	0.901	0.781	0.953	0.849	**0.718**
	0.2 for $d=3,\ldots,7$	0.829	0.888	0.928	0.978	0.996	0.990	0.993	0.962	0.904	0.981	0.938	0.869
0.6	0.5 for $d=3,4$	0.912	0.924	0.944	0.986	0.991	0.993	1.000	0.999	0.989	0.999	0.997	0.983
	0.5 for $d=3,6$	0.722	0.754	0.796	0.924	0.946	0.957	1.000	0.996	0.955	0.998	0.986	0.933
	0.5 for $d=3,8$	**0.641**	**0.679**	0.720	0.865	0.899	0.916	0.999	0.992	0.929	0.996	0.977	0.896
	0.2 for $d=3,\ldots,7$	0.770	0.795	0.826	0.943	0.958	0.966	1.000	0.997	0.971	0.998	0.991	0.956
0.9	0.5 for $d=3,4$	0.902	0.905	0.927	0.981	0.982	0.987	0.999	0.999	0.991	1.000	1.000	0.988
	0.5 for $d=3,6$	0.698	0.704	0.747	0.905	0.910	0.927	0.997	0.997	0.962	1.000	0.999	0.950
	0.5 for $d=3,8$	**0.613**	**0.619**	**0.660**	0.836	0.843	0.866	0.995	0.996	0.940	1.000	0.999	0.922
	0.2 for $d=3,\ldots,7$	0.750	0.755	0.788	0.929	0.933	0.945	0.998	0.998	0.976	1.000	1.000	0.967

Weights are: (a) $w_i=1$, (b) $w_i=1/(d_i-1)$, (c) $w_i=1/[(d_i-1)(1+0.5(d_i-1))]$ and (d) $w_i=1/[(d_i-1)d_i]$. Worst cases are in boldface.

This is different from the negative log-likelihood (applying algebra to (2.2)):

$$-L_0(\eta^2)=\sum_i\left\{\frac{1}{2}d_i\log\eta^2+\frac{1}{2}(d_i-1)\log(1-\rho)+\frac{1}{2}\log[1+(d_i-1)\rho]+\frac{\mathbf{z}_i'\mathbf{C}_i\mathbf{z}_i}{2\eta^2(1-\rho)}\right\},$$

where $\mathbf{C}_i=\mathbf{I}_{d_i}-\rho[1+(d_i-1)\rho]^{-1}\mathbf{J}_{d_i}$. Both $\mathbf{B}_i$ and $\mathbf{C}_i$ depend on ρ but not η^2 and μ. The BCL estimate of η^2 is

$$\widehat{\eta}^2_w=\frac{(1-\rho^2)^{-1}\sum_i w_i(\mathbf{y}_i-\mu\mathbf{1}_i)'\mathbf{B}_i(\mathbf{y}_i-\mu\mathbf{1}_i)}{\sum_i w_i d_i(d_i-1)}.$$

The MLE of η^2 is

$$\widehat{\eta}^2=\frac{(1-\rho)^{-1}\sum_i(\mathbf{y}_i-\mu\mathbf{1}_i)'\mathbf{C}_i(\mathbf{y}_i-\mu\mathbf{1}_i)}{\sum_i d_i}.$$

Unlike the estimation of μ with ρ known, in general no choice of weights will make the BCL estimate the same as the MLE.

Using results on moments of quadratic forms which are summarized in Appendix A.2, the variance of $\widehat{\eta}^2_w$ is

$$\mathrm{Var}\,(\widehat{\eta}^2_w)=(1-\rho^2)^{-2}\frac{\sum_i w_i^2\times 2\,\mathrm{tr}\,((\mathbf{B}_i\Sigma_i)^2)}{\left[\sum_i w_i d_i(d_i-1)\right]^2}=\frac{2\eta^4(1-\rho)^2}{(1-\rho^2)^2}\frac{\sum_i w_i^2 d_i(d_i-1)t_i}{\left[\sum_i w_i d_i(d_i-1)\right]^2},$$

where

$$t_i=d_i^2\rho^2+d_i(1+2\rho-3\rho^2)-(1+2\rho-3\rho^2).$$

The optimal choice of w_i is proportional to t_i^{-1}. Special cases are:

$$w_i\propto\begin{cases}(d_i-1)^{-1}, & \rho=0;\\ [0.25d_i^2+1.25d_i-1.25]^{-1}, & \rho=0.5;\\ d_i^{-2}, & \rho=1.\end{cases}$$

In the middle case of moderate dependence, based on a plot, the inverse weight for small d_i is close to $(d_i-1)[1+\frac{1}{2}(d_i-1)]$, which was considered earlier in (2.6). The other two case of inverse linear and quadratic weights occurred above.

The exchangeable multivariate normal model is simple enough to allow some analytic comparisons of weighted BCL to maximum likelihood. The various cases of estimating one parameter with others assumed known, suggest weights that are constant, or roughly inversely proportional to d_i or d_i^2. In Table 2, these different weights are compared when all three parameters μ, η^2, ρ are estimated simultaneously. Table 2 shows the patterns of the AREs with a few choices of ρ and distributions of cluster sizes; the discussion of this table is given jointly with that for Table 3. An outline of details behind the calculation of the Fisher information matrix and $\mathbf{V}$ is given in Appendix A.4.

Table 3
Multivariate exchangeable probit: ARE of BCL estimates of μ and ρ for different distributions of varying cluster sizes; $\mu = -0.84$, ρ chosen for weak, moderate and strong dependence; weights are: (a) $w_i = 1$, (b) $w_i = 1/(d_i - 1)$, (c) $w_i = 1/[(d_i - 1)(1 + 0.5(d_i - 1))]$ and (d) $w_i = 1/[(d_i - 1)d_i]$

ρ	Mixture	wt. in (a)		wt. in (b)		wt. in (c)		wt. in (d)	
		μ	ρ	μ	ρ	μ	ρ	μ	ρ
0.2	0.5 for $d = 3, 4$	0.949	0.986	0.998	0.978	0.995	0.941	0.990	0.927
	0.5 for $d = 3, 6$	0.830	0.953	0.988	0.939	0.974	0.770	0.948	0.720
	0.5 for $d = 3, 8$	0.780	0.927	0.974	0.920	0.951	0.641	0.909	**0.577**
	0.2 for $d = 3, \ldots, 7$	0.859	0.951	0.990	0.956	0.981	0.827	0.963	0.785
0.6	0.5 for $d = 3,4$	0.921	0.954	0.988	0.983	0.997	0.967	0.995	0.957
	0.5 for $d = 3, 6$	0.751	0.844	0.942	0.958	0.993	0.894	0.985	0.860
	0.5 for $d = 3, 8$	**0.676**	0.777	0.895	0.929	0.991	0.836	0.976	0.786
	0.2 for $d = 3, \ldots, 7$	0.791	0.850	0.953	0.951	0.993	0.911	0.987	0.886
0.9	0.5 for $d = 3, 4$	0.899	0.940	0.974	0.985	0.989	0.980	0.989	0.974
	0.5 for $d = 3, 6$	0.702	0.791	0.902	0.944	0.982	0.933	0.982	0.910
	0.5 for $d = 3, 8$	**0.617**	**0.709**	0.836	0.895	0.976	0.893	0.976	0.859
	0.2 for $d = 3, \ldots, 7$	0.747	0.804	0.919	0.936	0.979	0.934	0.979	0.918

Worst cases are in boldface.

2.3. Exchangeable probit model with varying cluster sizes

Multivariate probit models are more representative of where BCL is really needed in practice, and in this subsection, we report on some ARE analysis for weighted BCL for the exchangeable probit model with varying cluster sizes. Similar patterns obtain in comparison with the preceding subsection.

In Table 3, we summarize some AREs of the BCL estimates of the two parameters for four sets of weights, considered in the preceding subsection, with different distributions of cluster sizes. The three settings for ρ represent weak correlation, moderate correlation and strong correlation. The patterns are similar for different μ.

From Table 2 (for normal) and 3 (for probit), conclusions are the following.

1. The AREs of the BCL estimates decrease with larger cluster sizes and more variability in cluster sizes.
2. Constant weights (over cluster size) are not good, particularly for the parameter μ, but with these weights, the efficiency is not always worse than the weights in (b) for the dependence parameter; see (2.10).
3. The best choice of weight depends on the parameter and the strength of dependence.
   (i) With weak dependence, $w_i = 1/(d_i - 1)$ is best for all parameters, except $w_i = 1$ is best for ρ with the probit model.
   (ii) With moderate dependence, $w_i = 1/[(d_i - 1)(1 + 0.5(d_i - 1))]$ is best for μ (and η^2), $w_i = 1/(d_i - 1)$ is best for ρ, $w_i = 1/(d_i - 1)$ is marginally best overall.
   (iii) With strong dependence, $w_i = 1/[(d_i - 1)d_i]$ is best for μ (and η^2), $w_i = 1/(d_i - 1)$ or $1/[(d_i - 1)(1 + 0.5(d_i - 1))]$ is best for ρ, and $w_i = 1/[(d_i - 1)(1 + 0.5(d_i - 1))]$ is best overall.

In the two examples (exchangeable normal and exchangeable binary probit), the conclusions are similar. In Section 1, we have a simple explanation of why $w_i = 1/(d_i - 1)$ is natural for independence (and hence weak dependence) and why $w_i = 1/[(d_i - 1)d_i]$ is natural for perfect dependence (and hence strong dependence). The intermediate choice of $w_i = 1/[(d_i - 1)(1 + 0.5(d_i - 1))]$ does quite well over a range of moderate to strong dependence.

3. Longitudinal data

In this section, we do some analytic and numerical comparisons of AREs for constant length d for longitudinal series for n subjects. For multivariate normal, we assume the AR(1) covariance structure, and for binary probit, we assume the latent AR(1) correlation matrix. We do a comparison of ARE of BCL based on all pairs of bivariate margins, versus BCL based on the $(d - 1)$ bivariate margins with pairs of adjacent indices $(j, j + 1)$, $j = 1, \ldots, d - 1$. The latter is motivated on (1.2) and should be reasonable for models that are nearly Markovian; it is used in [25] and called a first-order pairwise likelihood. We could also consider the general weighting in (1.1) of the form $w_{i,jk} = w_{|j-k|}$ (weight depending on lag).

For non-constant lengths d_i, the discussion of weighting in Section 2 can give some insight, since the case of independence and perfect dependence can still be considered as boundary cases. If all bivariate pairs are used, then the discussion in Section 2 applies exactly. If only bivariate margins for adjacent pairs are used, then (a) for independent observations, the BCL is close to the log-likelihood with no weighting by cluster size (since the first and d_ith univariate margin would be counted once and the other univariate margins counted twice); (b) for perfect dependence, the BCL becomes the log-likelihood with a weight proportional to $(d_i - 1)^{-1}$ (since each of the $d_i - 1$ pairs provide the same information). A compromise is to use inverse weights by cluster size that are midway between the two boundary cases.

3.1. AR(1) normal model

In order to see some patterns, we study the AR(1) model $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{id})' \sim N(\mu \mathbf{1}_d, \eta^2 \mathbf{R}(\rho))$ where $\mathbf{R} = \mathbf{R}(\rho) = (\rho^{|j-k|})_{1\le j,k\le d}$ and $-1 < \rho < 1$. In this case, some analytic results are possible for the estimate of one of the parameters assuming the other two to be known. Let L_0 denote the log-likelihood, L_1 denote the BCL with all pairs, L_2 denote the BCL with only adjacent pairs, and L_3 denote the BCL with adjacent pairs but assuming the first and the last observations to be adjacent. With fixed cluster size d, then with the notation of (1.1), in L_1, $w_{i,jk} = 1$ for all i, j, k; in L_2, $w_{i,jk} = 1$ if $|j-k| = 1$ and 0 otherwise; and in L_3, $w_{i,jk} = 1$ if $|j-k| = 1$ or $|j-k| = d-1$ and 0 otherwise. One justification for the composite log-likelihood L_3 is given in subsection for AR(1) probit models. Another justification is that, by adding the $(1, d)$ bivariate margin to the adjacent bivariate margins, the resulting BCL becomes twice the log-likelihood at the boundary case of independence ($\rho = 0$), so that maybe it can be expected to do best under weak dependence.

Using results in Appendix A.1 together with the form of the bivariate normal density, with observation vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$, the log-likelihood and BCLs are:

$$-L_0 = \sum_{i=1}^{n} \left\{ c_0 \log \eta^2 + \frac{1}{2}(d-1)\log(1-\rho^2) + \frac{1}{2}\eta^{-2}(\mathbf{y}_i - \mu\mathbf{1}_d)'\mathbf{B}_0(\mathbf{y}_i - \mu\mathbf{1}_d) \right\}, \tag{3.1}$$

$$-L_1 = \sum_{i=1}^{n} \left\{ c_1 \log \eta^2 + \frac{1}{2}\sum_{k=1}^{d-1}(d-k)\log(1-\rho^{2k}) + \frac{1}{2}\eta^{-2}(\mathbf{y}_i - \mu\mathbf{1}_d)'\mathbf{B}_1(\mathbf{y}_i - \mu\mathbf{1}_d) \right\}, \tag{3.2}$$

$$-L_2 = \sum_{i=1}^{n} \left\{ c_2 \log \eta^2 + \frac{1}{2}(d-1)\log(1-\rho^2) + \frac{1}{2}\eta^{-2}(\mathbf{y}_i - \mu\mathbf{1}_d)'\mathbf{B}_2(\mathbf{y}_i - \mu\mathbf{1}_d) \right\}, \tag{3.3}$$

$$-L_3 = \sum_{i=1}^{n} \left\{ c_3 \log \eta^2 + \frac{1}{2}(d-1)\log(1-\rho^2) + \frac{1}{2}\log(1-\rho^{2(d-1)}) + \frac{1}{2}\eta^{-2}(\mathbf{y}_i - \mu\mathbf{1}_d)'\mathbf{B}_3(\mathbf{y}_i - \mu\mathbf{1}_d) \right\}, \tag{3.4}$$

where $c_0 = d/2$, $c_1 = d(d-1)/2$, $c_2 = d-1$ and $c_3 = d$, and $\mathbf{B}_0 = \mathbf{R}^{-1}$ (closed form given in Appendix A.1), $\mathbf{B}_2 = \mathbf{B}_0 + \text{diag}\,(0, \mathbf{1}'_{d-2}, 0)$,

$$\mathbf{B}_1 = \mathbf{B}_2 + \sum_{\ell=2}^{d-1}\sum_{j=1}^{d-\ell}(1-\rho^{2\ell})^{-1}\mathbf{A}_{j,j+\ell}, \qquad \mathbf{B}_3 = \mathbf{B}_2 + (1-\rho^{2(d-1)})^{-1}\mathbf{A}_{1,d},$$

and $\mathbf{A}_{j,j+\ell}$ has 1 in the (j,j), $(j+\ell, j+\ell)$ positions; $-\rho^\ell$ in the $(j, j+\ell)$, $(j+\ell, j)$ positions and zeros elsewhere. There is no simpler form for $\mathbf{B}_1$ for $d > 3$. Note that ρ, but neither μ or η^2, appears in the $\mathbf{B}$ matrices.

From (3.1) to (3.4), we have MLEs and BCL estimators

$$\widehat{\mu}_m = \sum_i \mathbf{1}'_d\mathbf{B}_m\mathbf{y}_i \Big/ \left[n\mathbf{1}'_d\mathbf{B}_m\mathbf{1}_d\right], \quad m = 0, 1, 2, 3; \tag{3.5}$$

and

$$\widehat{\eta}^2_m = \frac{\sum_i (\mathbf{y}_i - \mu\mathbf{1}_d)'\mathbf{B}_m(\mathbf{y}_i - \mu\mathbf{1}_d)}{2nc_m}, \quad m = 0, 1, 2, 3. \tag{3.6}$$

For estimation of ρ with μ, η^2 known: because $\mathbf{B}_2 - \mathbf{B}_0$ does not depend on ρ, the BCL estimate of ρ (based on $\mathbf{B}_2$) and the MLE of ρ are the same, and hence the BCL estimates of ρ with $\mathbf{B}_1$ and $\mathbf{B}_3$ are less efficient than that based on $\mathbf{B}_2$.

The variance of (3.5), with ρ assumed known, is

$$\text{Var}\,(\widehat{\mu}_m) = n^{-1}\eta^2\mathbf{1}'_d\mathbf{B}_m\mathbf{R}\mathbf{B}_m\mathbf{1}_d \Big/ \left[\mathbf{1}'_d\mathbf{B}_m\mathbf{1}_d\right]^2. \tag{3.7}$$

Using results in Appendix A.2 for moments of a quadratic form, the variance of (3.6), with μ, ρ known, is

$$\text{Var}\,(\widehat{\eta}^2_m) = \frac{1}{2}n^{-1}c_m^{-1}\eta^4\,\text{tr}\,\left([\mathbf{B}_m\mathbf{R}]^2\right). \tag{3.8}$$

Simplifications of these expressions for $m = 0, 2, 3$ are given in Appendix A.5.

When μ, η^2, ρ are simultaneously estimated, the outline of the details for computing $\mathbf{V}$ and the Fisher information matrix are given in Appendix A.4.

Table 4 reports some AREs, with d increasing for some selected values of ρ, for the three weighted BCL estimators. For μ, the AREs are the same whether η^2, ρ are estimated or assumed known (because of orthogonality in $\mathbf{V}$ and Fisher information). For η^2, AREs are included for the cases of μ, ρ (i) assumed known and (ii) simultaneously estimated with η^2.

Table 4
AR(1) normal model: ARE of weighted BCL for (1) all bivariate margins, (2) adjacent bivariate margins only, (3) adjacent bivariate margins and (1, d) margin

ρ	d	η^2, ρ estimated or not			μ, ρ known			μ, ρ estimated			μ, η^2 estimated		
		$\widehat{\mu}_1$	$\widehat{\mu}_2$	$\widehat{\mu}_3$	$\widehat{\eta}_1^2$	$\widehat{\eta}_2^2$	$\widehat{\eta}_3^2$	$\widehat{\eta}_1^2$	$\widehat{\eta}_2^2$	$\widehat{\eta}_3^2$	$\widehat{\rho}_1$	$\widehat{\rho}_2$	$\widehat{\rho}_3$
0.8	3	0.986	0.957	0.986	0.885	0.889	0.885	0.983	0.931	0.983	0.923	0.963	0.923
0.8	5	0.960	0.916	0.963	0.667	0.737	0.734	0.960	0.889	0.966	0.810	0.927	0.928
0.8	7	0.943	0.896	0.949	0.545	0.664	0.662	0.950	**0.879**	0.960	0.739	0.916	0.944
0.8	9	0.933	0.885	0.941	0.472	0.626	0.625	0.947	0.881	0.960	0.690	0.914	0.956
0.8	11	0.928	0.880	0.936	0.424	0.603	0.603	0.947	0.887	0.961	0.654	0.917	0.964
0.8	13	0.925	**0.878**	0.934	0.390	0.588	0.590	0.948	0.894	0.963	0.628	0.921	0.969
0.8	15	0.924	0.879	0.933	**0.365**	0.578	0.580	0.950	0.901	0.965	**0.608**	0.926	0.972
0.5	3	0.987	0.914	0.987	0.938	0.889	0.938	0.995	0.895	0.995	0.908	0.970	0.908
0.5	5	0.973	0.881	0.980	0.823	0.846	0.893	0.990	0.891	0.995	0.790	0.962	0.977
0.5	7	0.969	0.884	0.979	0.762	0.842	0.881	0.990	0.909	0.996	0.740	0.967	0.996
0.5	9	0.968	0.894	0.979	0.726	0.842	0.876	0.991	0.924	0.996	0.719	0.972	0.998
0.5	11	0.969	0.905	0.981	0.703	0.844	0.872	0.992	0.935	0.997	0.709	0.975	0.999
0.5	13	0.971	0.914	0.982	0.687	0.845	0.870	0.993	0.944	0.997	0.705	0.978	0.999
0.5	15	0.973	0.922	0.983	0.675	0.846	0.868	0.993	0.950	0.997	0.702	0.981	0.999
0.2	3	0.997	0.893	0.997	0.987	0.889	0.987	1.000	0.889	1.000	0.947	0.994	0.947
0.2	5	0.995	0.895	0.997	0.966	0.904	0.984	1.000	0.910	1.000	0.913	0.994	1.000
0.2	7	0.995	0.914	0.997	0.954	0.921	0.983	1.000	0.931	1.000	0.903	0.995	1.000
0.2	9	0.995	0.929	0.998	0.948	0.933	0.982	1.000	0.945	1.000	0.899	0.996	1.000
0.2	11	0.995	0.940	0.998	0.943	0.941	0.982	1.000	0.954	1.000	0.896	0.997	1.000
0.2	13	0.996	0.948	0.998	0.940	0.946	0.981	1.000	0.961	1.000	0.894	0.997	1.000
0.2	15	0.996	0.954	0.998	0.938	0.950	0.981	1.000	0.966	1.000	0.893	0.997	1.000
−0.5	3	0.997	0.914	0.997	0.938	0.889	0.938	0.995	0.895	0.995	0.908	0.970	0.908
−0.5	5	0.976	0.961	0.995	0.823	0.846	0.893	0.990	0.891	0.995	0.790	0.962	0.977
−0.5	10	0.969	0.983	0.996	0.713	0.843	0.874	0.991	0.930	0.996	0.713	0.974	0.999
−0.5	15	0.974	0.989	0.997	0.675	0.846	0.868	0.993	0.950	0.997	0.702	0.981	0.999
−0.8	3	1.000	0.957	1.000	0.885	0.889	0.885	0.983	0.931	0.983	0.923	0.963	0.923
−0.8	5	0.973	0.981	0.999	0.667	0.737	0.734	0.960	0.889	0.966	0.810	0.927	0.928
−0.8	10	0.936	0.995	0.994	0.445	0.613	0.613	0.946	0.883	0.960	0.671	0.915	0.960
−0.8	15	0.939	0.996	0.998	**0.365**	0.578	0.580	0.950	0.901	0.965	**0.608**	0.926	0.972

For ρ, the AREs are based on μ, η^2 estimated with ρ. When μ, η^2 are assumed known, the BCL estimator $\widehat{\rho}_2$ is the same as the MLE and so has full efficiency.

Patterns in $RE(\widehat{\mu}_m) = \text{Var}(\widehat{\mu}_{\text{MLE}})/\text{Var}(\widehat{\mu}_m)$ and $RE(\widehat{\eta}_m^2) = \text{Var}(\widehat{\eta}_{\text{MLE}}^2)/\text{Var}(\widehat{\eta}_m^2)$ are summarized next for the case of other parameters assumed known.

1. For fixed d, there is a constant $\rho_\mu(d) < 0$ such that $RE(\widehat{\mu}_1) \geq RE(\widehat{\mu}_2)$ for $\rho_\mu(d) < \rho < 1$. Also $RE(\widehat{\mu}_2) \leq RE(\widehat{\mu}_3)$ almost always except for some negative values of ρ. For fixed ρ, $RE(\widehat{\mu}_2) \to 1$ and $RE(\widehat{\mu}_3) \to 1$ as $d \to \infty$. Even when $\text{Var}(\widehat{\mu}_2)$ is largest, $RE(\widehat{\mu}_2)$ exceeds 0.877 (see derivation in Appendix A.4); $RE(\widehat{\mu}_3)$ has the same lower bound but the decrease towards it as d increases (with ρ changing) is much slower.
2. For fixed d, there are constants $\rho_\eta^-(d) < 0$ and $\rho_\eta^+(d) > 0$ such that $RE(\widehat{\eta}_2^2) \leq RE(\widehat{\eta}_1^2)$ for $\rho_\eta^-(d) < \rho < \rho_\eta^+(d)$. As d increases, the range where $\widehat{\eta}_1^2$ is better than $\widehat{\eta}_2^2$ becomes narrower. $RE(\widehat{\eta}_1^2)$ can be quite small for large d and $|\rho|$, whereas $RE(\widehat{\eta}_2^2)$ and $\text{RE}(\widehat{\eta}_3^2)$ do not get as small. For most (d, ρ) pairs, $RE(\widehat{\eta}_3^2)$ is the largest.
3. For fixed $\rho \neq 0$, with $d \geq d_\rho$ large enough, $RE(\widehat{\eta}_2^2) > RE(\widehat{\eta}_1^2)$.

In Table 4, a constant cluster size is being assumed, and unlike the exchangeable multivariate normal, the AREs are not 1 when all three parameters μ, η^2, ρ are simultaneously estimated. For η^2, the AREs of $\widehat{\eta}_m^2$ are larger when μ, ρ^2 are estimated (than when μ, ρ are assumed known) and are best for $\widehat{\eta}_3^2$. For ρ, the AREs are best for $\widehat{\rho}_3$, and the AREs for $\widehat{\rho}_1$ (all bivariate margins) can be low in cases of strong dependence. Overall, it is better to use all adjacent bivariate margins (plus $(1, d)$ margin for small d) than to use all bivariate margins.

3.2. AR(1) probit model

In this subsection, we study the AR(1) binary probit model with two parameters; there are similar patterns of AREs to Table 4. The AR(1) probit model can be used for longitudinal binary data; it is a model where composite likelihood methods are useful in practice to reduce the amount of computations (numerical integrals). Varin and Czado [22] use the more general autoregressive ordinal probit model.

The simplest AR(1) probit model without covariates has two parameters: the latent mean parameter μ and the latent correlation parameter ρ. For dimension d, the stochastic representation is:

$$Y_j = I(Z_j \leq \mu), \quad j = 1, \ldots, d, \qquad (Z_1, \ldots, Z_d)' \sim N(\mathbf{0}, \mathbf{R}(\rho)),$$

where $\mathbf{R}(\rho) = (\rho^{|j-k|})_{1\le j,k\le d}$, $-1 < \rho < 1$. With observed data $(y_{i1}, \ldots, y_{id})'$, $i = 1, \ldots, n$, we consider a weighted BCL of the form:

$$L = \sum_{\ell=1}^{d-1} w_\ell \sum_{j=1}^{d-\ell} L_{j,j+\ell}, \tag{3.9}$$

where the weight of the (j, k) bivariate margin depends on the lag $\ell = k - j$, and

$$L_{jk} = \sum_{s=0}^{1}\sum_{t=0}^{1} n_{st}^{(jk)} \log p_{jk}(s, t; \mu, \rho), \qquad p_{jk}(s, t; \mu, \rho) = \Pr(Y_j = s, Y_k = t; \mu, \rho),$$

$n_{st}^{(jk)}$ $(s, t = 0, 1)$ are the counts for the (j, k) bivariate margin.

Intuitively, we want to mainly use bivariate margins with lag 1 in order to reduce the amount of computation for larger d. The AR(1) probit model is not Markov (only the latent process is Markov). To match the notation in the preceding subsection, (3.9) becomes the BCL with all pairs, L_1, when $w_1 = w_2 = \cdots = w_{d-1} = 1$; the BCL with adjacent bivariate margins, L_2, when $w_1 = 1$ and $w_2 = \cdots = w_{d-1} = 0$; and the BCL with adjacent pairs including the (y_1, y_d) pair, L_3, when $w_1 = w_{d-1} = 1$, $w_2 = \cdots = w_{d-2} = 0$.

As a preliminary analysis on the choice of weights, for small values of d, we did a regression analysis of $\log \Pr(Y_1 = y_1, \ldots, Y_d = y_d; \mu, \rho)$ on $\log \Pr(Y_j = y_j, Y_k = y_k; \mu, \rho)$ (all $1 \le j \le k \le d$) with 'data' collected from inputs with different $(y_1, \ldots, y_d, \mu, \rho)$. These probabilities are rectangle probabilities and can be computed with the methods of Genz [3] or Joe [9] (e.g., R packages *mvtnorm* or *mprobit*). For $d = 3, 4, 5, 6$, the pattern is as follows: (i) the largest regression coefficients were for the adjacent pairs $\log \Pr(Y_1 = y_1, Y_2 = y_2)$ and $\log \Pr(Y_{d-1} = y_{d-1}, Y_d = y_d)$, (ii) the second largest coefficients were for the adjacent pairs $\log \Pr(Y_2 = y_2, Y_3 = y_3), \ldots, \log \Pr(Y_{d-2} = y_{d-2}, Y_{d-1} = y_{d-1})$ (iii) the third largest coefficient was for $\log \Pr(Y_1 = y_1, Y_d = y_d)$. The regression R^2 did not change much when other non-adjacent pairs were deleted. When log of univariate marginal probabilities $\log \Pr(Y_j = y_j)$ were included, the additional regression coefficients were close to zero.

In the actual computations of AREs of BCL estimators, it turned out that the choice of $w_1 = w_{d-1} = 1$ with other $w_\ell = 0$ (corresponding to L_3) is good. Table 5 has some representative results to show how the ARE behaves over different weight vectors $(w_1, \ldots, w_{d-1})$ and different d. The pattern is similar for different choices of (μ, ρ), so mainly one pair is used in Table 5 to make it easier to see patterns as d increases. The computations of Godambe matrices and AREs were done via the equations in Appendix A.6; the derivatives of the probabilities were computed using the methods in the R package *mprobit*.

The best choice of weight depends on whether the parameter μ or ρ is of primary interest. Similar to the multivariate normal AR(1) model, there are choices of the weights w_ℓ that do well without the need for all pairs.

1. For estimating ρ, the ARE can be better by decreasing w_l for $\ell > 1$.
2. For estimating μ, the ARE can be quite a bit worse when only adjacent bivariate margins are used compared with using all bivariate margins. In this case of $w_1 = 1$ and $w_\ell = 0$ for $\ell > 1$, the ARE of ρ is generally well over 0.9, but that for μ sometimes goes down to around 0.88.
3. If $w_1 = w_{d-1} = 1$ and other $w_\ell = 0$, then AREs for μ, ρ are well over 0.9.
4. If $w_1 = 0$, then the AREs for estimating ρ can be small.

The conclusions are similar for multivariate binary and normal. Unlike the case of clustered data in Section 2, the ARE of the BCL estimator with adjacent bivariate margins does not necessarily decrease as d increases.

3.3. Stochastic volatility model

The analyses in the preceding subsections are intended to provide some guidelines to the performance of weighted BCL for some stochastic volatility models for financial time series [6]. For a single time series with a latent AR(1) process, consider the AR(1) stochastic volatility model:

$$Y_t = \sigma_t \epsilon_t, \quad \log \sigma_t^2 = \alpha + \rho \log \sigma_{t-1}^2 + V_t, \tag{3.10}$$

with $-1 < \rho < 1$, where ϵ_t are independent N(0, 1) random variables, and V_t are independent N(0, σ_V^2) random variables, $\{\epsilon_t\}$ and $\{V_t\}$ are mutually independent, and α is a constant. The parameters $\xi = \alpha/(1-\rho)$ and $\omega = \sigma_V/\sqrt{1-\rho^2}$ are the stationary mean and standard deviation of $\{\log \sigma_t^2\}$.

Given observations $y_1, \ldots, y_n$, the joint density is

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \int \left\{ \prod_{j=1}^{n} s_j^{-1} f_\epsilon(y_j/s_j) \right\} f_{\sigma_1,\ldots,\sigma_n}(s_1, \ldots, s_n)\, \mathrm{d}s_1 \cdots \mathrm{d}s_n,$$

where $f_{\sigma_1,\ldots,\sigma_n}$ involves a density for the AR(1) process $\{\log \sigma_t^2\}$. The dimension of the integrand increases with n.

Bivariate marginal densities can easily be computed via Gauss-Hermite quadrature, so that weighted BCL is feasible. The asymptotic theory for BCL in this case is quite different from that for clustered data, as some ergodicity results are needed.

Table 5
AR(1) probit model: ARE of BCL for different weights by lag

d	μ	ρ	$w_1, \ldots, w_{d-1}$	ARE($\widehat{\mu}_w$)	ARE($\widehat{\rho}_w$)
3	−0.52	0.5	1.0, 1.0	0.996	0.965
3	−0.52	0.5	1.0, 0.0	0.912	0.946
3	−0.52	0.5	0.0, 1.0	0.877	0.357
3	−0.52	0.5	1.0, 0.8	0.990	0.983
3	−0.52	0.5	1.0, 0.5	0.976	0.997
3	−0.84	0.2	1.0, 1.0	0.999	0.982
3	−0.84	0.2	1.0, 0.0	0.893	0.984
4	−0.52	0.5	1.0, 1.0, 1.0	0.991	0.922
4	−0.52	0.5	1.0, 0.0, 0.0	0.899	0.925
4	−0.52	0.5	1.0, 0.0, 1.0	0.993	0.933
4	−0.52	0.5	0.0, 1.0, 0.0	0.987	0.436
4	−0.52	0.5	0.0, 0.0, 1.0	0.764	0.124
4	−0.52	0.5	1.0, 1.0, 0.0	0.945	0.952
4	−0.52	0.5	1.0, 0.5, 0.0	0.930	0.986
4	−0.52	0.5	1.0, 0.5, 0.5	0.972	0.986
4	−0.52	0.5	1.0, 0.0, 0.5	0.966	0.951
4	−0.84	0.5	1.0, 1.0, 1.0	0.992	0.921
4	−0.84	0.5	1.0, 0.0, 0.0	0.898	0.914
4	−0.84	0.5	1.0, 0.0, 1.0	0.994	0.927
5	−0.52	0.5	1.0, 1.0, 1.0, 1.0	0.988	0.887
5	−0.52	0.5	1.0, 0.0, 0.0, 0.0	0.897	0.914
5	−0.52	0.5	1.0, 0.0, 0.0, 1.0	0.991	0.919
5	−0.52	0.5	0.0, 1.0, 0.0, 0.0	0.911	0.474
5	−0.52	0.5	0.0, 0.0, 1.0, 0.0	0.933	0.169
5	−0.52	0.5	1.0, 1.0, 0.0, 0.0	0.912	0.948
5	−0.52	0.5	1.0, 0.0, 1.0, 0.0	0.956	0.920
5	−0.52	0.5	1.0, 1.0, 1.0, 0.0	0.957	0.907
5	−0.52	0.5	1.0, 1.0, 0.0, 1.0	0.971	0.941
5	−0.52	0.5	1.0, 0.0, 1.0, 1.0	0.991	0.889
5	−0.52	0.5	1.0, 0.5, 0.0, 0.0	0.908	0.981
5	−0.52	0.5	1.0, 0.0, 0.5, 0.0	0.939	0.948
5	−0.52	0.5	1.0, 0.0, 0.0, 0.5	0.963	0.925
5	−0.52	0.5	1.0, 0.5, 0.0, 0.5	0.954	0.983
5	−0.52	0.5	1.0, 0.5, 0.5, 0.5	0.972	0.973
6	−0.52	0.5	1.0, 1.0, 1.0, 1.0, 1.0	0.985	0.863
6	−0.52	0.5	1.0, 0.0, 0.0, 0.0, 0.0	0.900	0.908
6	−0.52	0.5	1.0, 0.0, 0.0, 0.0, 1.0	0.990	0.911
6	−0.52	0.5	1.0, 1.0, 0.0, 0.0, 0.0	0.900	0.947

Various estimation approaches have been proposed since the quasi-maximum likelihood method in [6], with the best being the simulated maximum likelihood method proposed in [20]; see this latter paper for a comparison of various methods. The model (3.10) is an alternative to ARCH/GARCH models for financial time series, when there are significant autocorrelations in $\{|y_t|\}$ or $\{y_t^2\}$. For financial time series data, the estimated ρ for (3.10) often exceeds 0.9.

For our Monte Carlo simulations, comparing various weighted BCL estimators with the simulated MLE (see http://www.doornik.com for implementation in Ox); we find patterns similar to those reported in Section 3.1. In particular, the ARE for the parameters ρ and α become small as ρ increases towards 1. The ARE of the parameters ξ is quite good, as it is related to the mean of $\{\log y_t^2\}$. The efficiency of BCL with $w_1 = 1$ and $w_\ell = 0$ for $\ell \geq 2$ is close to 1 for $\rho < 0.8$, but for $\rho > 0.8$ the efficiency of BCL is worse but improves with more lags such as using $w_2 = 1$ and $w_3 = 1$. The poor behavior of BCL is mainly due to heavy left-skewness of sampling distribution of $\widehat{\rho}$ when the true ρ increases towards 1, and adding more lags lessens the left-skewness. Details of the theory and simulations will appear in a separate article. However Table 6 has some BCL estimates for some recent financial stock return data to show the patterns we observed in simulations.

4. Discussion

We have shown how different weightings can be considered for BCL for (i) clustered data with varying cluster sizes and (ii) longitudinal data modeled with a (latent) autoregressive process. Through a combination of analytical and numerical examples, we study a number of factors that can affect the ARE, and suggest as a practically good choice the use of weight $w_i = (d_i - 1)^{-1}[1 + \frac{1}{2}(d_i - 1)]^{-1}$ in general for clustered data. With an underlying AR(1) process, we suggest that mainly adjacent (or near adjacent) pairs be used (so that rather than $d(d-1)/2$ terms, there are $O(d)$ terms, and this makes a big difference for large d).

Table 6
BCL estimates up to lag 3, and comparison with simulated maximum likelihood (ML); Google, Hewitt-Packard, Microsoft, Intel, Yahoo from 19 August 2004 to 29 April 2008

Stock	Method	σ_V	ρ	α	ξ	ω
goog	BCL (lag = 1)	0.85	0.48	0.57	1.09	0.97
goog	BCL (lags = 1, 2)	0.71	0.68	0.35	1.10	0.97
goog	BCL (lags = 1, 2, 3)	0.65	0.75	0.28	1.10	0.97
goog	ML/Ox	0.41	0.88	0.14	1.19	1.90
hwp	BCL (lag = 1)	0.69	0.68	0.23	0.71	0.94
hwp	BCL (lags = 1, 2)	0.75	0.61	0.27	0.70	0.94
hwp	BCL (lags = 1, 2, 3)	0.71	0.66	0.24	0.71	0.94
hwp	ML/Ox	0.68	0.68	0.23	0.71	1.27
msft	BCL (lag = 1)	0.91	0.43	−0.02	−0.04	1.01
msft	BCL (lags = 1, 2)	0.76	0.64	−0.01	−0.04	1.00
msft	BCL (lags = 1, 2, 3)	0.67	0.74	−0.01	−0.04	0.99
msft	ML/Ox	0.35	0.93	0.00	0.06	2.45
intc	BCL (lag = 1)	0.72	0.33	0.52	0.78	0.76
intc	BCL (lags = 1, 2)	0.60	0.61	0.30	0.78	0.76
intc	BCL (lags = 1, 2, 3)	0.58	0.64	0.28	0.78	0.76
intc	ML/Ox	0.29	0.90	0.09	0.85	1.50
yhoo	BCL (lag = 1)	1.00	0.24	0.78	1.02	1.03
yhoo	BCL (lags = 1, 2)	0.96	0.35	0.66	1.02	1.03
yhoo	BCL (lags = 1, 2, 3)	0.91	0.46	0.55	1.02	1.03
yhoo	ML/Ox	0.94	0.34	0.68	1.04	1.07

Time series are 100 times daily financial returns.

Situations where BCL might become less efficient are the following. (i) cluster sizes $d_i \to \infty$; (ii) choice of weight $w_i = (d_i - 1)^{-1}$ and large variability of cluster sizes (e.g., some clusters of size 2 and other large clusters); (iii) strong dependence with large cluster size (as in Table 4 in Section 3.1); (iv) parameters not identifiable from bivariate margins. For (iv), one should consider composite likelihood based on trivariate or high-dimensional marginal likelihoods in order that parameters are identified. Where unweighted BCL does poorly in efficiency, we found that in some cases that unweighted trivariate composite log-likelihood also does poorly.

Topics for future research include the study of optimal weights when weights are functions of parameters, used in the efficient method of moments [5,8], and the use of weighting functions that need not be the same for all parameters. For the latter, a system of nonlinear equations would be required so it is a little more difficult to implement compared with optimizing a weighted BCL.

Some of our results, such as in Section 3.1, on weighted BCL are a bit unexpected. For more complicated models where comparisons with maximum likelihood are not possible, and where weighting of bivariate margins is relevant, we recommend that comparisons of different versions of weighted BCL be made via Godambe information matrices and/or Monte Carlo simulations.

Composite likelihood methods could also be applied to generalized linear mixed models for the clustered data, but for this class of models, the Laplace approximation or h-likelihood methods might be feasible (see [12,13]).

Acknowledgments

This research has been supported by an NSERC Canada grant. We thank the referees and the associate editor for detailed valuable comments.

Appendix

A.1. Results for exchangeable and AR(1) multivariate normal

For the exchangeable d-variate normal distribution, with mean vector and covariance matrix given in (2.1):

1. $-1/(d-1) \le \rho \le 1$ in order that Σ is a non-negative definite.
2. $|\Sigma| = \eta^{2d}(1-\rho)^{d-1}[1+(d-1)\rho]$, $\Sigma^{-1} = [\eta^2(1-\rho)]^{-1}[\mathbf{I}_d + c_d \mathbf{J}_d]$, $c_d = -\rho/[1+(d-1)\rho]$.
3. From the previous item, the joint density when $-1 < \rho < 1$ is

$$(2\pi)^{-d/2}\left\{\eta^{2d}(1-\rho)^{d-1}[1+(d-1)\rho]\right\}^{-1/2}\exp\{-QF\},$$

where, with $y_+ = \sum_{j=1}^d y_j$, the quadratic form is

$$QF = \frac{1}{2\eta^2(1-\rho)}\left[(\mathbf{y}-\mu\mathbf{1}_d)^T(\mathbf{y}-\mu\mathbf{1}_d) - \frac{\rho}{[1+(d-1)\rho]}(y_+ - d\mu)^2\right].$$

For the AR(1) d-variate normal distribution, we use the following results (derivation follows directly from the likelihood) in Section 3. Let $(Z_1, \ldots, Z_d)' \sim N(\mathbf{0}, \mathbf{R})$ where $\mathbf{R} = (\rho^{|j-k|})_{1\le j,k\le d}$ and $-1 < \rho < 1$. The determinant is $|\mathbf{R}| = (1-\rho^2)^{d-1}$, and the inverse $\mathbf{R}^{-1} = (\rho^{(jk)})$ satisfies

$$\rho^{(11)} = \rho^{(dd)} = (1-\rho^2)^{-1}, \qquad \rho^{(jj)} = \frac{1+\rho^2}{1-\rho^2}, \quad j = 2, \ldots, d-1;$$
$$\rho^{(jk)} = \frac{-\rho}{1-\rho^2} \quad \text{if } |k-j| = 1, \qquad \rho^{(jk)} = 0 \quad \text{if } |k-j| > 1.$$

A.2. Results on moments of quadratic forms

We list some background results on moments of quadratic forms in normal random variables; see [18]. Let $\mathbf{Z} \sim N(\mathbf{0}, \Sigma)$ and A, B, C be square matrices with the dimension of Σ.

1. $\mathrm{E}(\mathbf{Z}'A\mathbf{Z}) = \mathrm{tr}(A\Sigma)$.
2. $\mathrm{Var}(\mathbf{Z}'A\mathbf{Z}) = 2\,\mathrm{tr}((A\Sigma)^2)$.
3. $\mathrm{Cov}(\mathbf{Z}'A\mathbf{Z}, \mathbf{Z}'B\mathbf{Z}) = 2\,\mathrm{tr}(A\Sigma B\Sigma)$.
4. $\mathrm{E}[C\mathbf{Z}\cdot\mathbf{Z}'B\mathbf{Z}] = \mathbf{0}$.

A.3. Some calculations for the case of estimating ρ with μ, η^2 known in exchangeable multivariate normal model

Let $g = n^{-1}(1-\rho^2)^2\partial L_1/\partial\rho$ be the estimating equation for ρ, where $\partial L_1/\partial\rho$ is the derivative of (2.9). Then

$$ng = -\frac{1}{2}\rho(1-\rho^2)\sum_i w_i d_i(d_i-1) + \rho\sum_i w_i(d_i-1)\mathbf{z}_i'\mathbf{z}_i - \frac{(1+\rho^2)}{2}\sum_i w_i\mathbf{z}_i'\mathbf{A}_i\mathbf{z}_i.$$

For the Godambe information, we need $D = \mathrm{E}[-\partial g/\partial\rho]$ and $M = n\mathrm{Var}(g)$. Straightforward calculations lead to

$$n\frac{\partial g}{\partial\rho} = -\frac{1}{2}(1-3\rho^2)\sum_i w_i d_i(d_i-1) + \sum_i w_i(d_i-1)\mathbf{z}_i'\mathbf{z}_i - \rho\sum_i w_i\mathbf{z}_i'\mathbf{A}_i\mathbf{z}_i$$

and

$$D = \mathrm{E}[-\partial g/\partial\rho] = -\frac{1}{2}(1+\rho^2)n^{-1}\sum_i w_i d_i(d_i-1). \tag{A.1}$$

With $\mathbf{R}_i = (1-\rho)\mathbf{I}_{d_i} + \rho\mathbf{J}_{d_i}$ and using the identities in Appendix A.2, let

$$a_{i1} = \mathrm{Var}(\mathbf{Z}_i'\mathbf{Z}_i) = 2\,\mathrm{tr}[(\mathbf{R}_i)^2] = 2d_i[1+(d_i-1)\rho^2],$$
$$a_{i2} = \mathrm{Var}(\mathbf{Z}_i'\mathbf{A}_i\mathbf{Z}_i) = 2\,\mathrm{tr}[(\mathbf{A}_i\mathbf{R}_i)^2] = 2d_i(d_i-1)[d_i^2\rho^2 + d_i(2\rho-3\rho^2) + 1 - 4\rho + 3\rho^2],$$
$$a_{i3} = \mathrm{Cov}(\mathbf{Z}_i'\mathbf{Z}_i, \mathbf{Z}_i'\mathbf{A}_i\mathbf{Z}_i) = 2\,\mathrm{tr}[\mathbf{R}_i\mathbf{A}_i\mathbf{R}_i] = 2d_i(d_i-1)\rho[2+(d_i-2)\rho].$$

Hence

$$nM = \rho^2\sum_i w_i^2(d_i-1)^2 a_{i1} + \frac{1}{4}(1+\rho^2)^2\sum_i w_i^2 a_{i2} - 2\rho\frac{(1+\rho^2)}{2}\sum_i w_i^2 a_{i3} = 2\sum_i w_i^2 b_i, \tag{A.2}$$

where, with the help of symbolic manipulation software,

$$b_i = \frac{1}{4}d_i(d_i-1)(1-\rho)^2\left[d_i^2\rho^2(1-\rho)^2 + d_i(2\rho-3\rho^2+8\rho^3-3\rho^4) + (1-\rho)^2(1+3\rho^2)\right].$$

A.4. Derivatives and Godambe matrix for general form of weighted BCL for multivariate normal models

A general form that covers various weighted BCL for exchangeable and AR(1) normal models, for a fixed dimension d, is:

$$L_w = -\lambda(\rho) - c\log\eta^2 - \frac{1}{2}\eta^{-2}(\mathbf{y}-\mu\mathbf{1})'\mathbf{K}(\rho)(\mathbf{y}-\mu\mathbf{1}),$$

where $\lambda(\rho)$ are terms with the log determinant, $c = c(d)$, and $\mathbf{K}(\rho)$ is a $d \times d$ matrix. The first order partial derivatives are:

$$\frac{\partial L_w}{\partial \mu} = \eta^{-2}\mathbf{1}'\mathbf{K}(\rho)(\mathbf{y} - \mu\mathbf{1}),$$
$$\frac{\partial L_w}{\partial \eta^2} = -c(d)\eta^{-2} + \frac{1}{2}\eta^{-4}(\mathbf{y} - \mu\mathbf{1})'\mathbf{K}(\rho)(\mathbf{y} - \mu\mathbf{1}),$$
$$\frac{\partial L_w}{\partial \rho} = -\lambda'(\rho) - \frac{1}{2}\eta^{-2}(\mathbf{y} - \mu\mathbf{1})'\frac{\partial \mathbf{K}(\rho)}{\partial \rho}(\mathbf{y} - \mu\mathbf{1}).$$

The $\mathbf{M}_d$ matrix comes from the covariance of the above first order derivatives, and $\mathbf{D}_d$ matrix has the expected values of the negative Hessian of L_w. The entries of the $\mathbf{D}_d$ and $\mathbf{M}_d$ matrices have entries as shown in the following table.

Element	$\mathbf{D}_d$	$\mathbf{M}_d$
(μ, μ)	$\eta^{-2}\mathbf{1}'\mathbf{K}(\rho)\mathbf{1}$	$\eta^{-2}\mathbf{1}'\mathbf{K}(\rho)\mathbf{R}(\rho)\mathbf{K}(\rho)\mathbf{1}$
(μ, η^2)	0	0
(μ, ρ)	0	0
(η^2, η^2)	$-c(d)\eta^{-4} + \eta^{-4}\,\mathrm{tr}\,(\mathbf{K}(\rho)\mathbf{R}(\rho))$	$\frac{1}{2}\eta^{-4}\,\mathrm{tr}\,(\mathbf{K}(\rho)\mathbf{R}(\rho)\mathbf{K}(\rho)\mathbf{R}(\rho))$
(η^2, ρ)	$-\frac{1}{2}\eta^{-2}\,\mathrm{tr}\left(\frac{\partial \mathbf{K}(\rho)}{\partial \rho}\mathbf{R}(\rho)\right)$	$-\frac{1}{2}\eta^{-2}\,\mathrm{tr}\left(\mathbf{K}(\rho)\mathbf{R}(\rho)\frac{\partial \mathbf{K}(\rho)}{\partial \rho}\mathbf{R}(\rho)\right)$
(ρ, ρ)	$\lambda''(\rho) + \frac{1}{2}\,\mathrm{tr}\left(\frac{\partial^2 \mathbf{K}(\rho)}{\partial \rho^2}\mathbf{R}(\rho)\right)$	$\frac{1}{2}\,\mathrm{tr}\left(\frac{\partial \mathbf{K}(\rho)}{\partial \rho}\mathbf{R}(\rho)\frac{\partial \mathbf{K}(\rho)}{\partial \rho}\mathbf{R}(\rho)\right)$

For a fixed cluster size, the inverse Godambe matrix for the BCL estimate is $\mathbf{V} = \mathbf{D}_d^{-1}\mathbf{M}_d\mathbf{D}_d^{-1}$. If there are varying cluster sizes such that dimension d has probability π_d and the weight for cluster size d is w_d, then the asymptotic covariance matrix of the BCL estimator is

$$\left(\sum_d \pi_d w_d \mathbf{D}_d\right)^{-1}\left(\sum_d \pi_d w_d^2 \mathbf{M}_d\right)\left(\sum_d \pi_d w_d \mathbf{D}_d\right)^{-1}.$$

A.5. Relative efficiency analysis of weighted BCL for AR(1) normal model

When just μ or η^2 is estimated assuming other parameters are known, some analytic results can be obtained. For relative efficiency comparisons, we substitute in closed forms for the matrix expressions in (3.7) and (3.8) if possible. The details are a bit tedious so mainly we show the final expressions which have been checked numerically for correctness.

(i) $\widehat{\mu}_0$: $\mathbf{1}_d'\mathbf{B}_0\mathbf{1}_d = \mathbf{1}_d'\mathbf{R}^{-1}\mathbf{1}_d = (1-\rho^2)^{-1}[d + (d-2)\rho^2 - 2(d-1)\rho] = (1+\rho)^{-1}[d - (d-2)\rho]$ and $n\eta^{-2}\mathrm{Var}\,(\widehat{\mu}_0) = (\mathbf{1}_d'\mathbf{R}^{-1}\mathbf{1}_d)^{-1}$.

(ii) $\widehat{\mu}_2$: $\mathbf{1}_d'\mathbf{B}_2\mathbf{1}_d = \mathbf{1}_d'\mathbf{B}_0\mathbf{1}_d + (d-2) = (1-\rho^2)^{-1}[2d - 2 - 2(d-1)\rho] = 2(d-1)(1+\rho)^{-1}$, $\mathbf{1}_d'\mathbf{B}_2\mathbf{R}\mathbf{B}_2\mathbf{1}_d = \mathbf{1}_d'\mathbf{R}^{-1}\mathbf{1}_d + 2(d-2) + \sum_{\ell=0}^{d-3}(d-2-\ell)\rho^\ell$

$$= \frac{d-(d-2)\rho}{1+\rho} + (d-2) + 2\frac{(d-2)-(d-1)\rho+\rho^{d-1}}{(1-\rho)^2}.$$

Then

$$n\eta^{-2}\mathrm{Var}\,(\widehat{\mu}_2) = \frac{\mathbf{1}_d'\mathbf{B}_2\mathbf{R}\mathbf{B}_2\mathbf{1}_d}{(\mathbf{1}_d'\mathbf{B}_2\mathbf{1}_d)^2} = \frac{(1+\rho)(2d-3-2d\rho+\rho+\rho^{d-1}+\rho^d)}{2(d-1)^2(1-\rho)^2}.$$

From (i), the relative efficiency function is

$$RE_2(d,\rho) = \frac{\mathrm{Var}\,(\widehat{\mu}_0)}{\mathrm{Var}\,(\widehat{\mu}_2)} = \frac{2(d-1)^2(1-\rho)^2}{[d-(d-2)\rho]\cdot[2d-3-2d\rho+\rho+\rho^{d-1}+\rho^d]}. \tag{A.3}$$

For (A.3), let $\widetilde{\rho}(d)$ be the argmin of $RE_2(d,\rho)$ for fixed d. For example, $\widetilde{\rho}(d)$ is 0, 0.4753, 0.7378, 0.8241 respectively and $RE_2(d,\widetilde{\rho}(d))$ is 0.8889, 0.8811, 0.8784, 0.8780 respectively for $d = 3, 5, 10, 15$. $RE_2(d,\widetilde{\rho}(d))$ is slowly decreasing in d, and $\widetilde{\rho}(d) \to 1$ as $d \to \infty$. But also $RE_2(d,\rho) \to 1$ as $d \to \infty$ for any fixed ρ. To further analyze (A.3), substitute $1-\rho = a/d$ where $a = \lim_{d\to\infty} d(1-\widetilde{\rho}(d))$. Then (A.3) becomes

$$\frac{2a^2}{[2+a-2ad^{-1}][ad^{-1}(2d-1)-2+(1-ad^{-1})^d(2-ad^{-1})/(1-ad^{-1})]} \sim \frac{a^2}{(2+a)(a-1+e^{-a})}. \tag{A.4}$$

For $a > 0$, the right-hand side of (A.4) is minimized at $a = 2.6880$ with value 0.87769 for the minimum relative efficiency.

(iii) $\widehat{\mu}_3$: Similarly, with the help of symbolic manipulation software,

$$n\eta^{-2}\mathrm{Var}\,(\widehat{\mu}_3) = \frac{\mathbf{1}_d'\mathbf{B}_3\mathbf{R}\mathbf{B}_3\mathbf{1}_d}{(\mathbf{1}_d'\mathbf{B}_3\mathbf{1}_d)^2} = \frac{(1+\rho)\,C}{2(1-\rho)^2(d\rho+\rho^2+(d-1)\rho^d)^2(\rho+\rho^d)},$$

where

$$\begin{aligned} C = {} & 2d\rho^3(1-\rho) - 3\rho^5 + \rho^6 + (6d-4)\rho^{d+2} + (2-6d)\rho^{d+3} - 4\rho^{d+4} + 2\rho^{d+5} + (6d-7)\rho^{2d+1} \\ & + (5-6d)\rho^{2d+2} + \rho^{2d+3} + \rho^{2d+4} + (2d-2)\rho^{3d} + (4-2d)\rho^{3d+1} + 2\rho^{3d+2} + \rho^{4d-1} + \rho^{4d} \end{aligned}$$

and

$$RE_3(d,\rho) = \frac{\mathrm{Var}\,(\widehat{\mu}_0)}{\mathrm{Var}\,(\widehat{\mu}_3)} = \frac{2(1-\rho)^2(d\rho+\rho^2+(d-1)\rho^d)^2(\rho+\rho^d)}{[d-(d-2)\rho]\,C}. \tag{A.5}$$

For (A.5), let $\widetilde{\rho}_3(d)$ be the argmin of $RE_3(d,\rho)$ for fixed d. For example, $\widetilde{\rho}_3(d)$ is 0.6791, 0.7716, 0.8517, 0.8864 respectively and $RE_3(d,\widetilde{\rho}_3(d))$ is 0.9837, 0.9624, 0.9352, 0.9215 respectively for $d = 3, 5, 10, 15$. $RE_3(d,\widetilde{\rho}_3(d))$ is slowly decreasing in d, and $\widetilde{\rho}_3(d) \to 1$ as $d \to \infty$. The limit as $d \to \infty$ of the $RE_3(d,\widetilde{\rho}_3(d))$ is the same as $RE_2(d,\widetilde{\rho}_3(d))$ since as $d \to \infty$, the effect of the $(1,d)$ margin goes to 0. That is, the $(1,d)$ margin helps for efficiency only for small d.

(iv) $\widehat{\mu}_1$: $\mathbf{1}_d'\mathbf{B}_1\mathbf{1}_d = \mathbf{1}_d'\mathbf{B}_2\mathbf{1}_d + \sum_{\ell=2}^{d-1}(d-\ell)(1-\rho^{2\ell})^{-1}(2-2\rho^{\ell})$, $\mathbf{1}_d'\mathbf{B}_1\mathbf{R}\mathbf{B}_1\mathbf{1}_d$ does not simplify but can be evaluated numerically.

(v) $\widehat{\eta}_2^2, \widehat{\eta}_3^2$: With some algebraic simplifications, $n\mathrm{Var}\,(\widehat{\eta}_0^2) = 2\eta^4 d^{-1}$, and

$$n\eta^{-4}\mathrm{Var}\,(\widehat{\eta}_2^2) = 0.5c_2^{-2}\,\mathrm{tr}\,\left((\mathbf{B}_2\mathbf{R})^2\right) = \frac{(d-1)(1-3\rho^2+\rho^4)+d-2+\rho^{2(d-1)}}{(d-1)^2(1-\rho^2)^2}.$$

As $d \to \infty$, this behaves like $(2-\rho^2)/[d(1-\rho^2)]$. Hence as $d \to \infty$,

$$RE(\widehat{\eta}_2^2) \to \frac{2(1-\rho^2)}{2-\rho^2} = \frac{(1-\rho^2)}{(1-\rho^2/2)}.$$

This is 1 for $\rho = 0$ and 0 as $\rho \to \pm 1$. The limit is between 0 and 1 otherwise, and decreases as $|\rho|$ increases; it is the same for $RE(\widehat{\eta}_3^2)$.

(vi) $\widehat{\eta}_1^2$: The derivation of $n\eta^{-4}\mathrm{Var}\,(\widehat{\eta}_1^2)$ is not tractable. Numerically it has been checked that $\lim_{d\to\infty} RE(\widehat{\eta}_1^2) \le 1-\rho^2 \le (1-\rho^2)/(1-\rho^2/2)$.

A.6. Outline of probability calculations needed for Godambe matrix for multivariate discrete distribution

Suppose we have a d-dimensional discrete distribution for $\Pr(Y_1 = y_1, \ldots, Y_d = y_d)$ with parameter $\boldsymbol{\theta}$. From a sample of size n, let $n_{st}^{(jk)}$ be the observations/counts for the (j,k) bivariate margin, and let $p_{jk}(s,t;\boldsymbol{\theta})$ be the (j,k) bivariate probability mass function. The weighted BCL is:

$$L_w = \sum_{1\le j<k\le d} w_{jk} \sum_{s,t} n_{st}^{(jk)} \log p_{jk}(s,t;\boldsymbol{\theta}).$$

Under regularity conditions, the BCL estimator $\tilde{\boldsymbol{\theta}}$ is the root of the equation

$$\frac{\partial L_w}{\partial \boldsymbol{\theta}} = \sum_{j<k} w_{jk} \sum_{s,t} n_{st}^{(jk)} \mathbf{h}_{jk}(s,t;\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j<k} w_{jk} \sum_{s,t} I(y_{ij}=s, y_{ik}=t)\,\mathbf{h}_{jk}(s,t;\boldsymbol{\theta}),$$

where $\mathbf{h}_{jk}(s,t;\boldsymbol{\theta}) = \partial \log p_{jk}/\partial\boldsymbol{\theta}$. The relevant estimating function is

$$\mathbf{g} = \sum_{1\le j<k\le d} w_{jk} \sum_{s,t} I(y_j=s, y_k=t)\,\mathbf{h}_{jk}(s,t;\boldsymbol{\theta}).$$

The asymptotic covariance matrix, as $n \to \infty$, for $\tilde{\boldsymbol{\theta}}$ is $\mathbf{V} = \mathbf{D}^{-1}\mathbf{M}(\mathbf{D}')^{-1}$ where $\mathbf{M} = \mathrm{E}\,(\mathbf{g}\,\mathbf{g}')$ and $\mathbf{D} = \mathrm{E}\,[-\partial\mathbf{g}/\partial\boldsymbol{\theta}']$. Note that

$$\mathbf{D} = \sum_{j<k} w_{jk} \sum_{s,t} \frac{\partial p_{jk}}{\partial\boldsymbol{\theta}} \frac{\partial p_{jk}}{\partial\boldsymbol{\theta}'}/p_{jk} = \sum_{j<k} \mathcal{J}_{jk},$$

where $\mathcal{J}_{jk}$ is the Fisher information matrix from estimation of $\boldsymbol{\theta}$ from the (j,k) bivariate margin. If not all of the components of $\boldsymbol{\theta}$ appear in the bivariate margin, then there are rows and columns of zeros in $\mathcal{J}_{jk}$.

Next $\mathbf{g}\mathbf{g}' = \sum_{j<k} w_{jk} \sum_{j^*<k^*} w_{j^*k^*} \sum_{s,t}\sum_{s^*,t^*} I(y_j=s, y_k=t)\,I(y_{j^*}=s^*, y_{k^*}=t^*)\,\mathbf{h}_{jk}(s,t;\boldsymbol{\theta})\,\mathbf{h}_{j^*k^*}'(s^*,t^*;\boldsymbol{\theta})$, so $\mathbf{M} = \sum_{j<k} w_{jk} \sum_{j^*<k^*} w_{j^*k^*} \sum_{s,t}\sum_{s^*,t^*} \Pr(Y_j=s, Y_k=t, Y_{j^*}=s^*, Y_{k^*}=t^*)\,\mathbf{h}_{jk}(s,t;\boldsymbol{\theta})\,\mathbf{h}_{j^*k^*}'(s^*,t^*;\boldsymbol{\theta})$. The double sum over (j,k) and (j^*,k^*) can be divided into three cases: number of distinct elements in j,k,j^*,k^* is 2, 1 or 0. For these cases, the inner sums become:

(i) $\sum_{s,t} p_{jk}(s,t;\boldsymbol{\theta})\,\mathbf{h}_{jk}(s,t;\boldsymbol{\theta})\,\mathbf{h}'_{jk}(s,t;\boldsymbol{\theta})$;
(ii) $\sum_{s,t,t^*} p_{jkk^*}(s,t,t^*;\boldsymbol{\theta})\,\mathbf{h}_{jk}(s,t;\boldsymbol{\theta})\,\mathbf{h}'_{jk^*}(s,t^*;\boldsymbol{\theta})$ with pairs (j,k) and (j,k^*), $k\neq k^*$;
(iii) $\sum_{s,t,s^*,t^*} p_{jkj^*k^*}(s,t,s^*,t^*;\boldsymbol{\theta})\,\mathbf{h}_{jk}(s,t;\boldsymbol{\theta})\,\mathbf{h}'_{j^*k^*}(s^*,t^*;\boldsymbol{\theta})$,

where p_{jkk^*}, $p_{jkj^*k^*}$ are trivariate and 4-variate marginal distributions.

For dimension d, there are $\binom{d}{2}$ terms of type (i), $d(d-1)(d-2)$ terms of type (ii), and $d(d-1)(d-2)(d-3)/4$ terms of type (iii). Note that the calculations for $\mathbf{V}$ depends just on the bivariate, trivariate and 4-variate marginal probabilities. If some weights are zero, then not all terms are needed.

References

[1] D.R. Cox, N. Reid, A note on pseudolikelihood constructed from marginal densities, Biometrika 91 (2004) 729–737.
[2] A.R. de Leon, Pairwise likelihood approach to grouped continuous model and its extension, Statist. Probab. Lett. 75 (2005) 49–57.
[3] A. Genz, Numerical computation of multivariate normal probabilities, J. Comput. Graph. Statist. 1 (1992) 141–149.
[4] H. Geys, G. Molenberghs, S. Lipsitz, A note on the comparison of pseudo-likelihood and generalized estimating equations for marginally specified odds ratio models with exchangeable association structure, J. Stat. Comput. Simul. 62 (1998) 45–71.
[5] L.P. Hansen, Large sample properties of generalized method of moments estimators, Econometrica 50 (1982) 1029–1054.
[6] A.C. Harvey, E. Ruiz, N. Shephard, Multivariate stochastic variance models, Rev. Econom. Stud. 61 (1994) 247–264.
[7] P.J. Heagerty, S.R. Lele, A composite likelihood approach to binary spatial data, J. Amer. Statist. Assoc. 93 (1998) 1099–1111.
[8] G.W. Imbens, One-step estimators for over-identified generalized method of moments models, Rev. Econom. Stud. 64 (1997) 359–383.
[9] H. Joe, Approximations to multivariate normal rectangle probabilities based on conditional expectations, J. Amer. Statist. Assoc. 90 (1995) 957–964.
[10] A.Y.C. Kuk, D.J. Nott, A pairwise likelihood approach to analyzing correlated binary data, Statist. Probab. Lett. 47 (2000) 319–335.
[11] S. Le Cessie, J.C. Van Houwelingen, Logistic regression for correlated binary data, Appl. Statist. 43 (1994) 95–108.
[12] Y. Lee, J.A. Nelder, Hierarchical generalized linear models, J. R. Statist. Soc. B 54 (1996) 3–40 (with Discussion).
[13] Y. Lee, J.A Nelder, Y. Pawitan, Generalized Linear Models with Random Effects, Chapman & Hall/CRC, Boca Raton, FL, 2006.
[14] S. Lele, M.L. Taper, A composite likelihood approach to (co)variance components estimation, J. Statist. Plann. Inference 103 (2002) 117–135.
[15] B. Lindsay, Composite likelihood methods, in: N.U. Probhu (Ed.), Statistical Inference from Stochastic Processes, American Mathematical Society, Providence, RI, 1988, pp. 221–239.
[16] A. Maydeu-Olivares, H. Joe, Limited information goodness-of-fit testing in multidimensional contingency tables, Psychometrika 71 (2006) 713–732.
[17] D.J. Nott, T. Ryden, Pairwise likelihood methods for inference in image models, Biometrika 86 (1999) 661–676.
[18] C.R. Rao, Linear Statistical Inference and its Applications, Wiley, New York, 1973.
[19] D. Renard, G. Molenberghs, H. Geys, A pairwise likelihood approach to estimation in multilevel probit models, Comput. Statist. Data Anal. 44 (2004) 649–667.
[20] G. Sandmann, S.J. Koopman, Estimation of stochastic volatility models via Monte Carlo maximum likelihood, J. Econometrics 87 (1998) 271–301.
[21] C. Varin, On composite marginal likelihoods, Adv. Stat. Anal. 92 (2008) 1–28.
[22] C. Varin, C. Czado, A mixed probit model for the analysis of pain severity diaries, (2008) (submitted for publication).
[23] C. Varin, G. Host, Ø. Skare, Pairwise likelihood inference in spatial generalized linear mixed models, Biometrika 49 (2005) 1173–1191.
[24] C. Varin, P. Vidoni, A note on composite likelihood inference and model selection, Comput. Statist. Data Anal. 92 (2005) 519–528.
[25] C. Varin, P. Vidoni, Pairwise likelihood inference for ordinal categorical time series, Comput. Statist. Data Anal. 51 (2006) 2365–2373.
[26] Y. Zhao, H Joe, Composite likelihood estimation in multivariate data analysis, Canad. J. Statist. 33 (2005) 335–356.